Barcoding Strategies for Multiplexing of Samples Using a Long-Read Sequencing Technology

Walter Lee, Tyson Bowen, John Harting, David Hsu, Emilia Mollova, Eun Lee, Regina Lam, Lance Hepler, Brett Bowman, Michael Brown, Simeon Pinaula, Phil Jiao, David Alexander, Erik Miller
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Introduction

- We have developed barcoding reagents and workflows for multiplexing amplicons or fragmented native genomic (DNA) prior to Single-Molecule, Real Time (SMRT®) Sequencing (Figure 1).
- The long reads of PacBio's SMRT Sequencing enable detection of linked mutations across multiple kilobases (kb) of sequence.
- This feature is particularly useful in the context of mutational analysis or SNP confirmation, where a large number of samples are generated routinely.
- To validate this workflow, a set of 384 1.7-kb amplicons, each derived from variants of the Phi29 DNA polymerase gene, were barcoded during amplification, pooled, and sequenced on a single SMRT Cell.
- To demonstrate the applicability of the method to longer inserts, a library of 96 5-kb clones derived from the *E. coli* genome was sequenced.

Materials and Methods

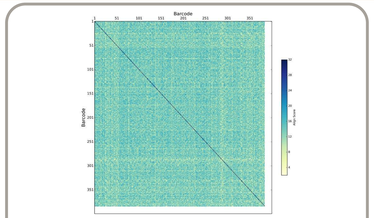

Figure 1. Barcodes for SMRT Sequencing. A set of 384 16-base barcodes was optimized for detection and discrimination in SMRT Sequencing. Pairwise alignment scores for the 384 barcodes are shown.

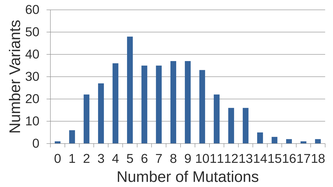

Figure 2. Variants generated from Phi29 mutagenesis.

- Two amplicon libraries were used in this demonstration:
  - A set of 384 variants of the Phi29 DNA polymerase gene, generated via Error-prone PCR of the 1.7-kb gene, containing 0-18 mutations (mean 7.2) relative to the input template (Figure 2).
  - A set of 96 unique 5-kb genomic regions cloned from *E. coli*.

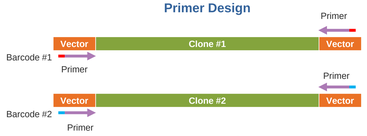

Figure 3. Barcoded samples were generated by amplification of targets with barcoded primers using Phusion® II polymerase (Thermo Scientific).

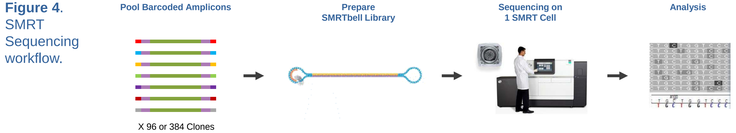

Figure 4. SMRT Sequencing workflow.

- Barcoded amplified products were pooled, purified and made into SMRTbell™ libraries using 10–15 µg of pooled barcoded amplicons according to standard protocols.
- Libraries were sequenced with P4-C2 chemistry for 90 minutes (Phi29 variants) or 180 minutes (*E. coli*).
- Sequencing data was clustered by barcode to generate multi-molecule consensus sequences for each construct present in the pool using pbbarcode and Long Amplicon Analysis (LAA) protocols as implemented in SMRT Analysis version 2.2. (Note: Clustering and phasing algorithms were turned off in LAA, resulting in a per-barcode *de novo* consensus sequence using subreads of minimum read quality 75 and read length 1500 bp for Phi29 and 4000 bp for *E. coli*).
- Consensus sequences in multi-FASTA format were compared against the corresponding consensus reference sequence derived from Sanger sequencing and Genbank, respectively, for concordance using BLASR (Pacific Biosciences).

Results

- SMRT Sequencing data were 100% concordant with independent Sanger sequencing, for a 100% accurate reconstruction of the set of clones.
- An error rate of 10^{-5} was obtained with 45X coverage, and no errors were detected above 50X coverage.

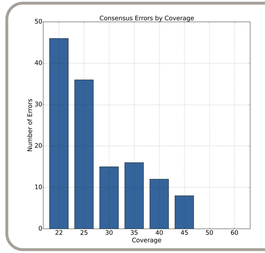

Figure 5. Coverage requirements at various error rates for assembly from a 700-kb subset of Phi29 sequence data.

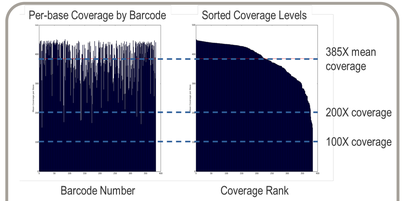

Figure 6. Coverage levels by barcode for 384 Phi29 variants. Simple pooling of PCR products resulted in >100X coverage for all clones in the dataset, while only 50X is required for accuracy of QV>50.

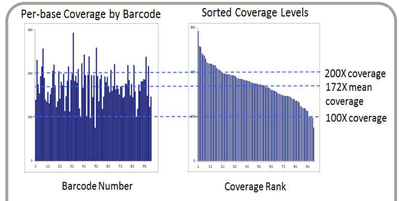

Figure 7. Coverage levels by barcode for 5-kb *E. coli* templates. Simple pooling of PCR products resulted in >75X coverage for all clones in the dataset, while only 50X coverage is required for accuracy of QV>50.

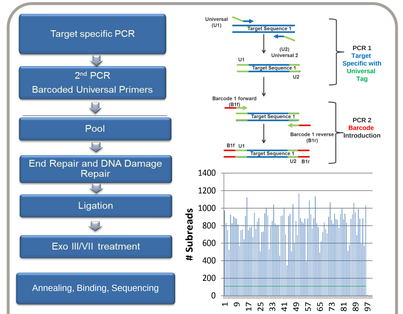

Figure 8. Amplicon multiplexing workflow for barcoded universal primer 96 multiplex kit.

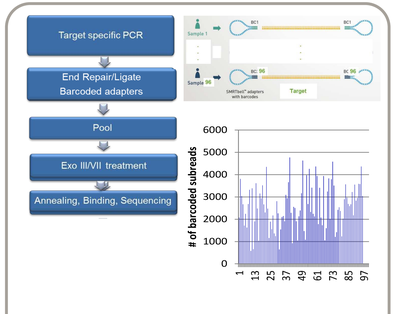

Figure 9. Alternate multiplexing workflow for existing amplicon workflows using an approach that integrates the barcode onto the SMRTbell adapter.

- In Figure 8, a two-step amplification utilizes target-specific primers that incorporate universal bridging sequences, followed by a second amplification with barcoded primers that also contain the bridging sequences. Following two-step PCR, amplicons were pooled and standard SMRTbell™ libraries were prepared. After Long Amplicon Analysis filtering, the yield of subreads per individual barcode far exceeded the 100 subreads required to generate consensus.
- In Figure 9, an alternate approach that incorporates the barcode into the SMRTbell adapter is demonstrated with an end-repair and ligation step. This method accommodates existing amplicon workflows.

Conclusions

- Long reads and high consensus accuracy make the PacBio® RS II an excellent system for high-throughput clone validation in protein engineering and display pipelines.
- Full-length reads allow easy sorting of highly similar sequences and confident linkage of multiple mutations within individual clones.
- Using barcoding, we demonstrate efficient and accurate single SMRT Cell 384 multiplex for 1.7-kb amplicons and 96 multiplex for 5-kb amplicons.
- Workflows have been established to incorporate >96 barcodes through either PCR or ligation, enabling pooling of samples from multiple sources.
- The low per-run cost and rapid turnaround of this method make it an excellent sequencing tool for selection pipelines.
- Products are available for 96-plex workflows using the Barcoded Universal Primer or Barcoded Adapter methodologies. PacBio has made available up to 384 barcodes and they can obtained at https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Barcoding